Editing Large Language Models: Problems, Methods, and Opportunities

Year: 2023 | *Citations: 386* | *Authors: Yunzhi Yao, Peng Wang, Bo Tian*

Abstract

Despite the ability to train capable LLMs, the methodology for maintaining their relevancy and rectifying errors remains elusive. To this end, the past few years have witnessed a surge in techniques for editing LLMs, the objective of which is to efficiently alter the behavior of LLMs within a specific domain without negatively impacting performance across other inputs. This paper embarks on a deep exploration of the problems, methods, and opportunities related to model editing for LLMs. In particular, we provide an exhaustive overview of the task definition and challenges associated with model editing, along with an in-depth empirical analysis of the most progressive methods currently at our disposal. We also build a new benchmark dataset to facilitate a more robust evaluation and pinpoint enduring issues intrinsic to existing techniques. Our objective is to provide valuable insights into the effectiveness and feasibility of each editing technique, thereby assisting the community in making informed decisions on the selection of the most appropriate method for a specific task or context. Code and datasets are available at https://github.com/zjunlp/EasyEdit.